Predictive Power of Diagnostics: Diabetes

Cristina Brackeen

Agenda

- Background Information
- Research Question
- Exploring the Data
- Optimizing Models
- Final Test
- Outcome
- Conclusion

Diabetes in America

- Approximately 30.3 million Americans have Diabetes
- Diabetes is the 7th leading cause of death in the United States
- A recent study indicates that regardless of reason for admission, Diabetic patients with severe dysglycemia are at increased risk for hospital readmission
- Additionally, the most common reason for readmission is heart failure, which has an extremely high death rate of 30% in 3 years

The Data

- UCI Repository on Diabetes 130-US Hospital from 1999-2008
  - Acquired from Health Facts Database which is a volunteer service provided to hospitals using Cerner Electronic Health Record System
- Data meet the following criteria extracted from Hospital databases:
  - Hospital admission
  - Diabetic encounter
  - Length of stay was between 1-14 days
  - Laboratory tests were performed
  - Medications were administered during encounter

Research Question

What factors predict whether diabetic patients will need to be readmitted for a future diabetic encounter?

Variables

Variable	Encounter ID	Patient No	Race	Gender	Age
Type	Integer	Integer	Categorical	Categorical	Categorical
NaNs	No	No	Yes	No	No
N	101766	101766	99493	101766	101766

Weight (lbs)	Admission type	Discharge disposition	Admission Source	Time in hospital (days)	Payer Code
Integer	Categorical	Categorical	Categorical	Integer	Categorical
Yes	No	No	No	No	Yes
3197	101766	101766	101766	101766	61510

Medical Specialty	Lab Procedures	Procedures	Medications	Out-paitient visits	Emergency visits
Categorical	Integer	Integer	Integer	Integer	Integer
No	No	No	No	No	No
51817	101766	101766	101766	101766	101766

Variables

Variable	Inpatient visits	Diagnosis 1	Diagnosis 2	Diagnosis 3	Diagnoses No
Type	Integer	Categorical	Categorical	Categorical	Integer
NaNs	No	Yes	Yes	Yes	No
N	101766	101745	101408	100343	101766

Glucose-serum	A1c	Change of Medication	Diabetic Medication	Readmitted
Categorical	Categorical	Binary	Binary	Categorical
No	No	No	No	No
101766	101766	101766	101766	101766

Variables

24 Categorical Medication features (Up, Down, Steady, None)
N = 101766

Medication	
Metformin	Acarbose
Repaglinide	Miglitol
Nateglinide	Troglitazone
Chlorpropamide	Tolazamide
Glimepiride	Examide
Acetohexamide	Sitagliptin
Glipizide	Insulin
Glyburide	Glyburide-metformin
Tolbutamide	Glipizide-metformin
Piogliztazone	Glimepiride-pioglitazone
Rosiglitazone	Metformin-Pioglitazone
Metformin-Rosiglitazone	

Initial Model

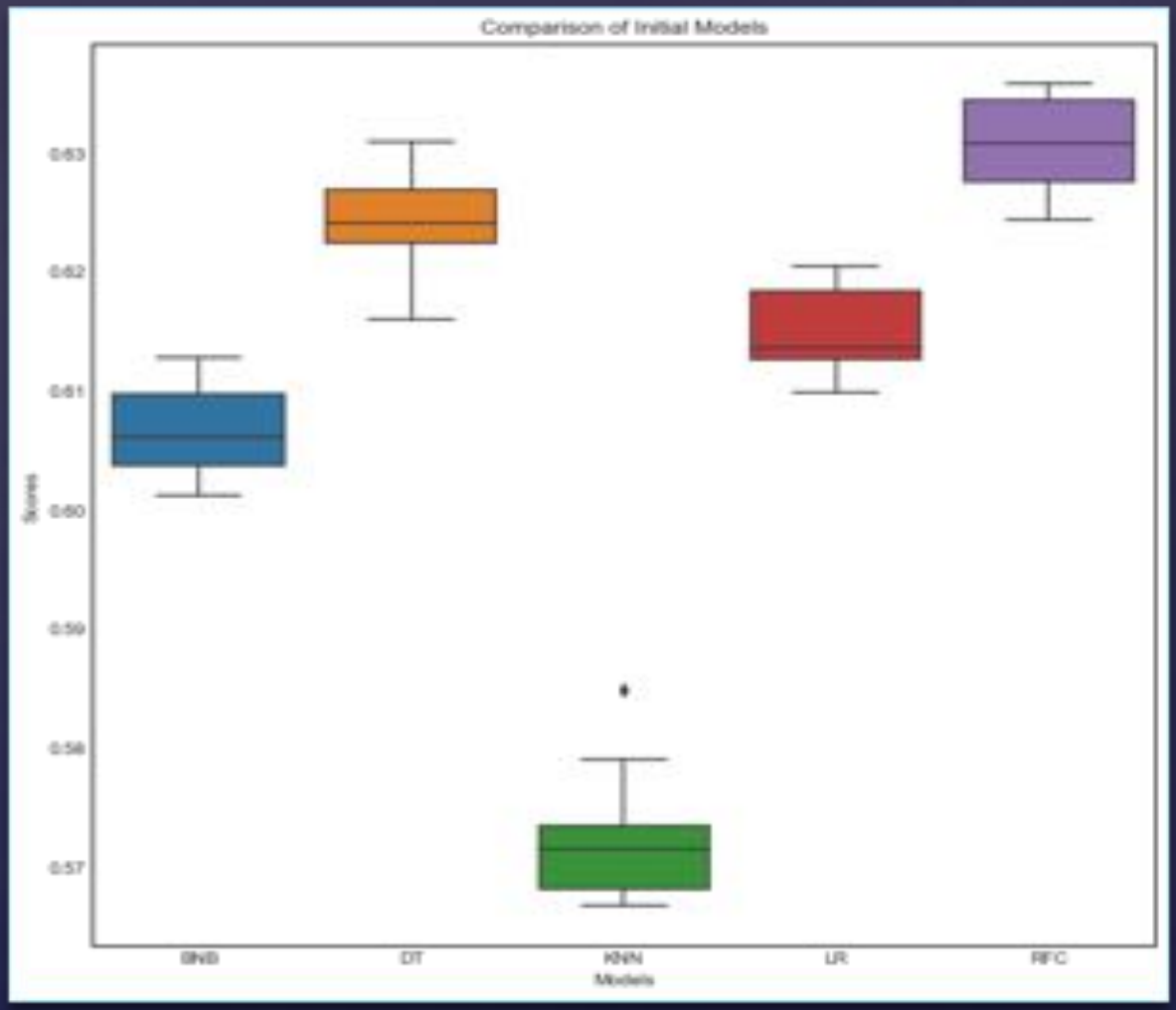
- N = 98,052
  - Dropped observations with NaNs
- Features = 91
  - Dropped features missing more than half of data *(weight, payer code, medical specialty)*
  - Dropped features that are simply identifying information *(encounter id, patient number, admission source id)*
  - Dropped categorical features with over 700 unique categories (Diagnosis 1, 2, 3)
- Outcome = Readmitted

Initial Model

- Observation = 98052, Predictors = 91
- Applied seven classifier methods:
  - Random Forest Classifier (RFC)
  - Decision Tree Classifier (DT)
  - Logistic Regression (LR)
  - K-Nearest Neighbor Classifier (KNN)
  - Naïve Bayes (BNB)
  - Support Vector Machines Classifier (SVM)
- Used 10 k-folds cross-validation scores with mean accuracy (standard deviation) (see above)

```
BNB : 0.607(0.004)
DT  : 0.624(0.004)
KNN : 0.572(0.005)
LR  : 0.615(0.003)
RFC : 0.631(0.004)
```

Initial Model

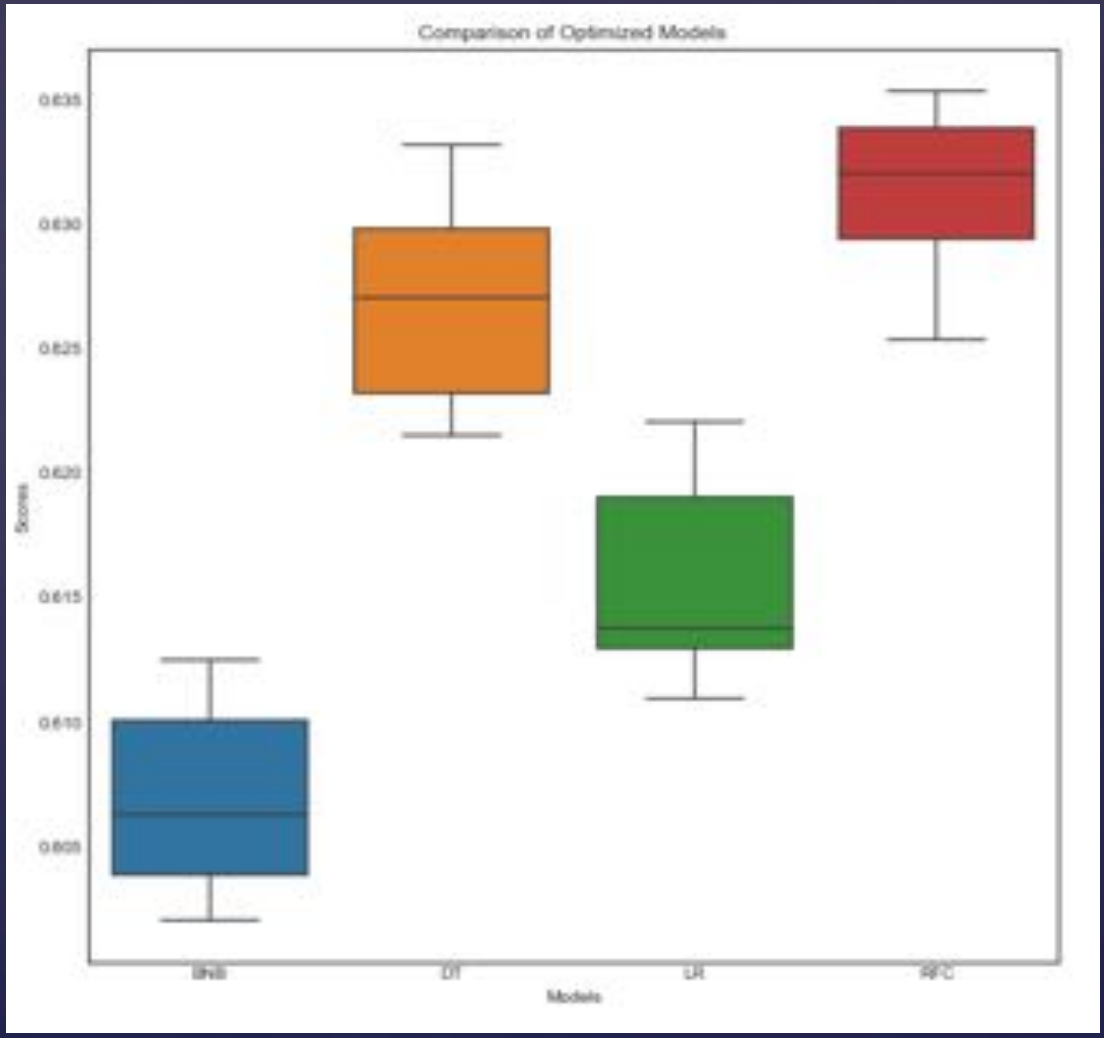


Optimized Parameter Model

- Observation = 98052, Predictors = 91
- Dropped KNN model due to poor performance
- Used GridSearch with Cross-Validation
  - Random Forest Classifier
    - Criterion = entropy
    - Max Depth = 10
    - N-estimators = 80
  - Naïve Bayes
    - Alpha = 50
  - Decision Tree Classifier
    - Criterion = entropy
    - Max Depth = 5
  - Logistic Regression
    - C = 0.01
- Use Cross-Validation Score with mean (standard deviation)

```
BNB : 0.607(0.004)
DT  : 0.627(0.004)
LR  : 0.616(0.004)
RFC : 0.631(0.003)
```

Optimized Parameter Model

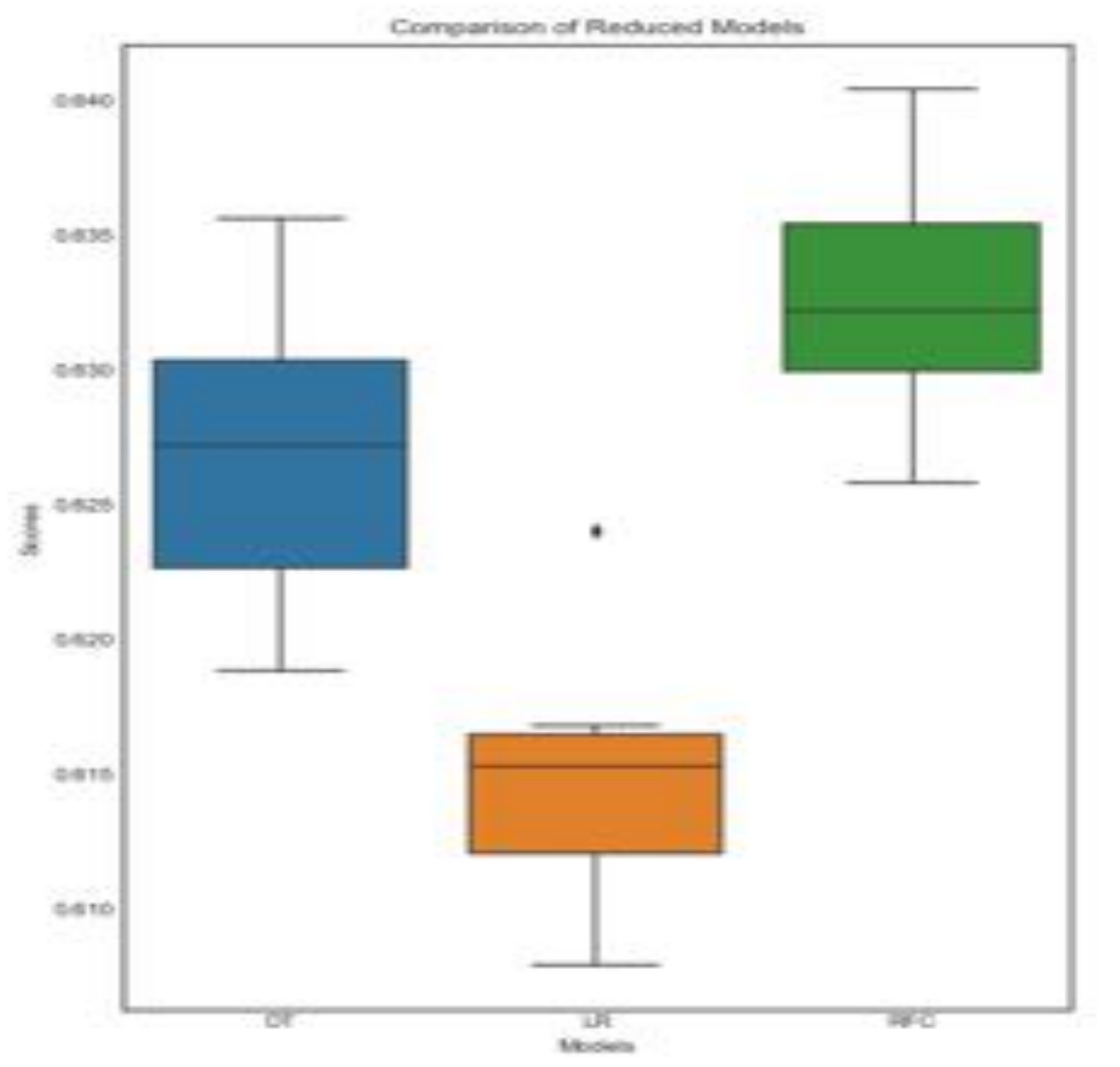


Reduced Features Model

- Observations: 98052, Predictors: 24
- Dropped BNB due to poor performance
- Used Recursive Feature Elimination with Cross-Validation
  - Ranked importance to model
  - Only took Rank = 1 Predictors
- Applied remaining three predictors
  - Random Forest Classifier
  - Decision Tree Classifier
  - Logistic Regression
- 10 k-fold Cross Validation with mean accuracy (standard deviation)

```
DT  : 0.627(0.005)
LR  : 0.615(0.004)
RFC : 0.633(0.005)
```

Reduced Features Model

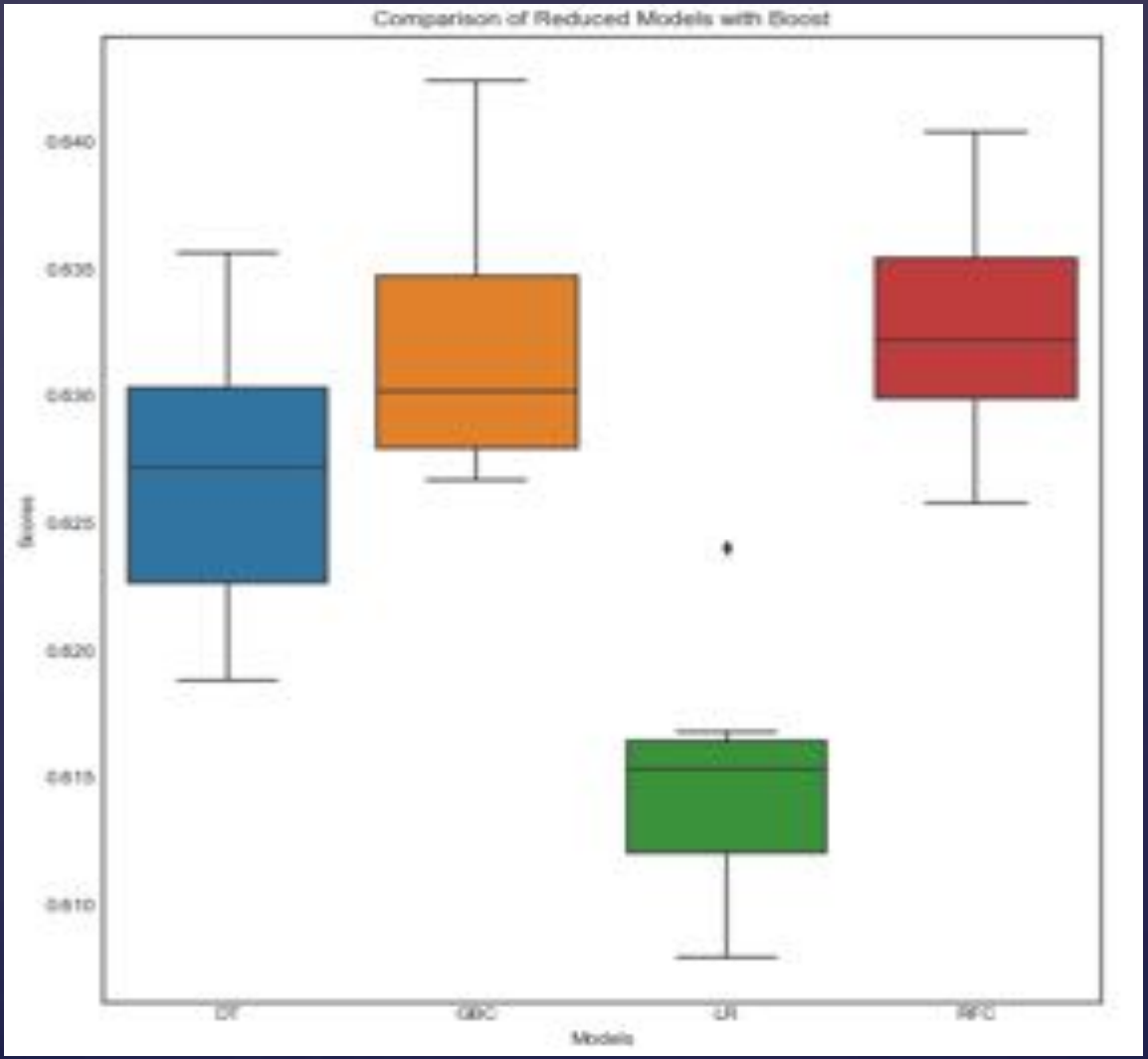


Boosting Model

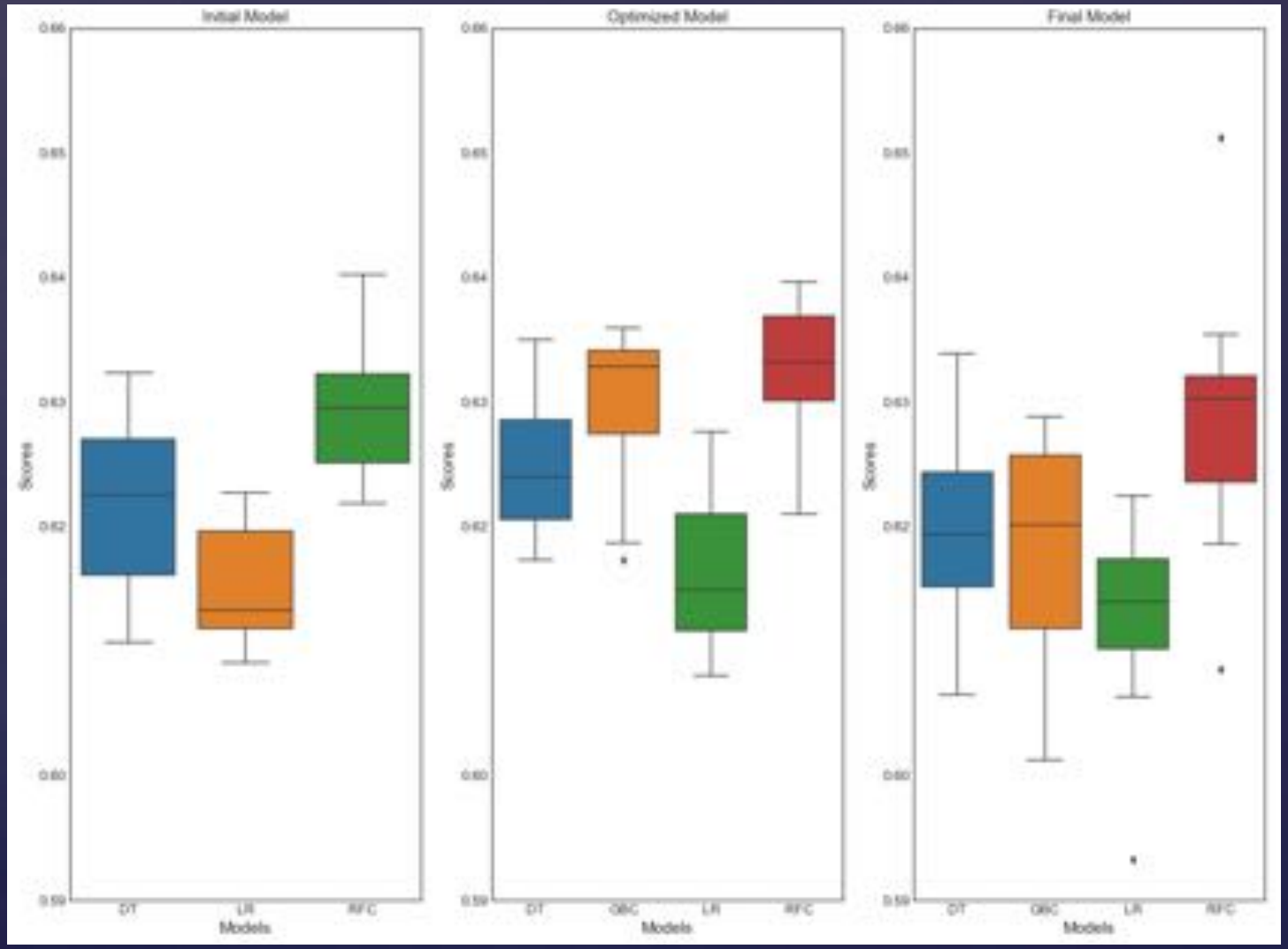
- Gradient Boosting Classifier
- Optimized using GridSearch
  - Loss = exponential
  - N-Estimators = 20
  - Max Depth = 10
- 10 k-fold Cross Validation with mean accuracy (standard deviation)

```
DT  : 0.627(0.005)
GBC : 0.632(0.005)
LR  : 0.615(0.004)
RFC : 0.633(0.005)
```

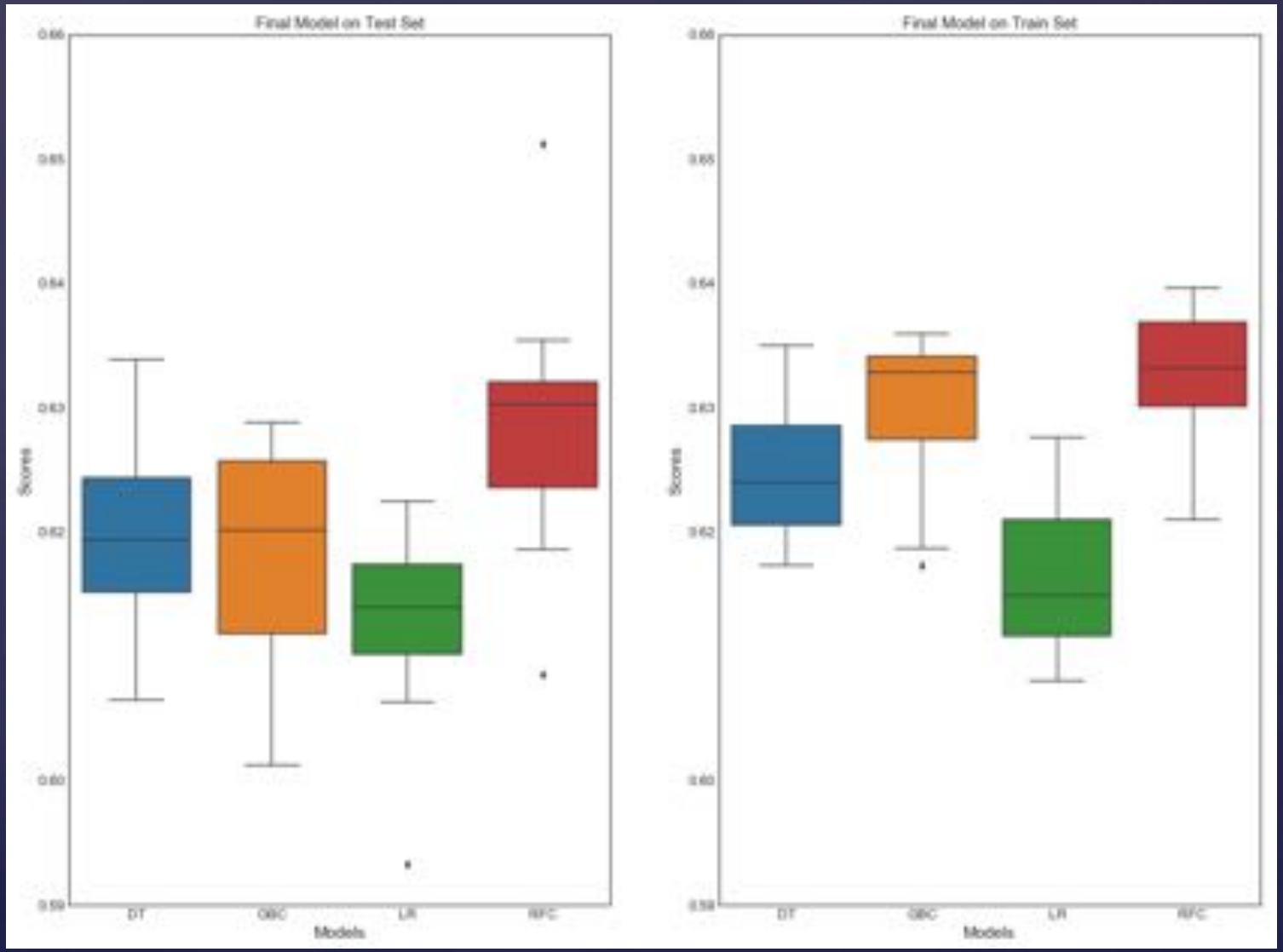
Boosting Model



Comparison of Initial, Optimized, and Final Models



Comparison of Test Group and Full Dataset



Outcome

- The main factors that help predict readmittance are: Gender, Age, Admission Type Id, Discharge Disposition Id, Time in Hospital, Number of Procedures done (lab and non-lab), Number of Medications administered, Number of hospital visits in the past year (Outpatient, Inpatient, and Emergency), Number of diagnoses entered, Change of Medications, Diabetes medications prescribed, Race(African American, Caucasian), Insulin (Down, No), Whether A1C and Glucose serum tests were taken, Whether Metformin was prescribed.

Concerns

- Causality
  - Are you more likely to have future inpatient stays if your past year was higher?
- Missing Information
  - The number one item that a lot of people associate with the disease: weight, had to be left out as 97% of the data was missing. Would this make a difference?
- Correlated Data
  - Some drugs are derivatives of others, i.e. metformin and glyburide-metformin, could this have an impact?
- Potential Bias
  - Data set came from Health Facts Database which is a volunteer service to hospitals using Cerner Electronic Health Record System

Further Research

- Compare to updated information in hospitals in the United States
  - Run Final Model to see if there are similar results
  - Optimize new data to see if predictors change over time

- Revisit the important predictors and see if we can lower readmittance rate
  - Running A1C tests or Glucose serum tests on all diabetic patients.
  - Revisit different Diabetes Medications given – prescribing Meds increases odds of

Conclusion

- The highest odds for readmittance are relating to number of inpatient stays (past year) and if diabetic medication was prescribed and can predict whether a patient could return for additional treatment
- We should be taking a look at what medications we are providing and in what doses to limit return inpatient visits as much as possible.
- These findings may or may not be representative for other countries

Questions?

Predictive Power of Diagnosis: Diabetes

https://github.com/cbrackeen/Diabetic-Readmittance

References

- https://rd.springer.com/chapter/10.1007%2F978-3-540-25966-4_33
- https://www.hindawi.com/journals/bmri/2014/781670/tab1/
- http://www.diabetes.org/diabetes-basics/statistics/
- https://www.healthleadersmedia.com/clinical-care/diabetes-complications-increase-readmission-risk